Aroofa Maknojia

About

World Health Organization Life Expectancy Machine Learning Model

Aroofa Maknojia Just now · 8 min read

Source: Everyday Health

Governments in individual countries work to improve their economy, infrastructure, military, and health care to create a better quality of life for their citizens. As developed nations continue to grow, they are widening the gap between themselves and the lower-

income developing countries that are struggling to develop their economies and general standard of living. Health and well-being are some of the major concerns that the World Health Organization (WHO) is tackling through the research they conduct and policies they promote. WHO keeps track of the progress countries have been making through the Global Health Observatory data repository they maintain. I explored a dataset of health, economic, and social factors for 193 countries between 2000–2015 from WHO.

The dataset had 22 columns and 2938 rows. After cleaning the dataset and removing null values, I was left with 1649 rows to work with. I also removed columns that were irrelevant to my analysis. The following columns are what I used for my analysis:

1. Status — (Developing or Developed)

2. Life Expectancy — (in age)

3. Adult Mortality — (rates of both sexes probability of dying between 15 and 60 years per 1000 population)

4. Infant deaths — (number of infant deaths per 1000 population)

5. Alcohol — (alcohol consumption recorded per capita in liters of pure alcohol)

6. Percentage expenditure -(expenditure on health as a percentage of domestic product per capita)

7. Hepatitis B — (HepB) immunization coverage among 1-year-olds (%)

8. Measles — Number of reported cases per 1000 population

9. BMI — Average body mass index of the entire population

10. Under-five deaths — Number of under-five deaths per 1000 population

11. Polio — Polio (Pol3) immunization coverage among 1-year-olds

12. Total expenditure — General government expenditure on health as a percentage of total government expenditure (%)

13. Diphteria — Diptheria tetanus toxoid and pertussis (DTP3) immunization covered among 1-year-olds (%)

14. HIV/AIDS — Deaths per 1000 live births HIV/AIDS (0–4 years)

15. GDP — Gross Domestic Product per capita (in USD)

16. Population — Population of the country

17. Thinness 1–19 years — Prevalence of thinness among children and adolescents for Ages 10–19 (%)

18. Thinness 5–9 years — Prevalence of thinness among children for Ages 5–9 (%)

19. Income composition — Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

20. Schooling — Number of years of schooling (years)

Here is the link to the code I use for my analysis and model creation.

Data Exploration

To get a better understanding of the range of data available, I found the minimum and maximum values of all of the variables. The lowest life expectancy for a country in the dataset is 44 while the highest is 89. There is a huge gap in the GDP for countries with the lowest GDP at 1.68 and the highest at 119,173. Access to education plays a big role worldwide with the lowest schooling level at 4.2 years and the highest going up to 20.7 years. These results make it obvious that some countries are developing at faster rates than other countries are and the largest gap is an economic one.

One of the columns in the dataset called Status classifies countries into 2 categories. Either the country is developing or it's developed. The count plot below shows the dataset has more representation of developing countries than developed countries. However, it is important to note that there are more developing countries in the world than developed countries, so the dataset is a good reflection of the world we live in. Status will be one of the predicting categorical variables for our machine learning model.

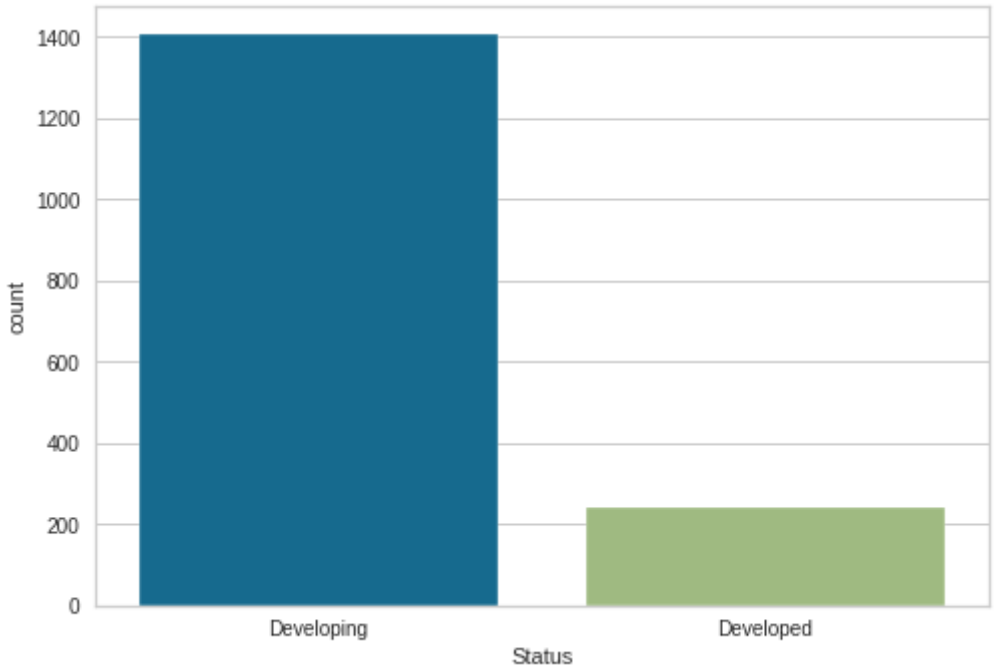

Figure 1: Count plot of Countries' Status (Developing/Developed)

I created a correlation matrix with all of the variables in the dataset to see if any of the variables are highly correlated with one another. The column I took out was the status column because it was a categorical variable while the others were numerical variables.

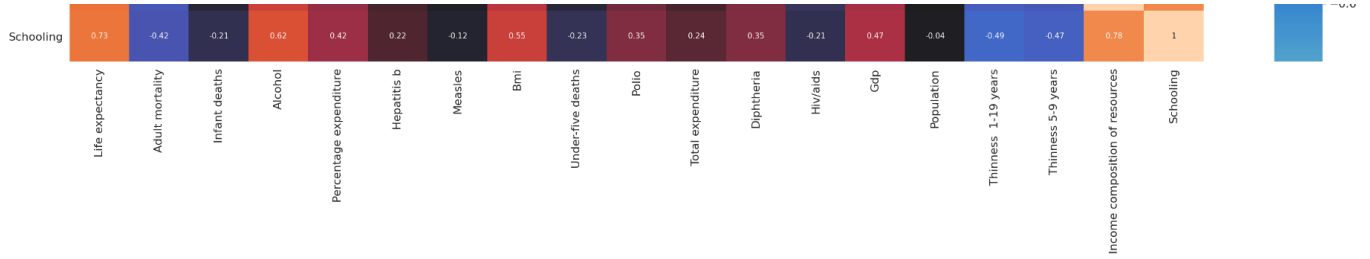

Figure 2: Correlation Matrix for All of the Variables

GDP and Percentage expenditure were two variables that showed a strong positive correlation with a correlation value of 0.96. Thinness 1–19 years and Thinness 5–9 years were also variables that showed a strong positive correlation with a correlation value of 0.93. On the other hand, adult mortality and life expectancy showed a strong negative correlation with a correlation value of -0.7. These correlations align with how we see developing and developed countries. Developed countries often have higher GDPs and households tend to spend more money on items to satisfy their daily needs and desires. Developing countries that have children that are underweight will have this issue for children of all age groups, not just one range. In a country where adults are passing away at higher rates, they will not have a larger life expectancy in their country.

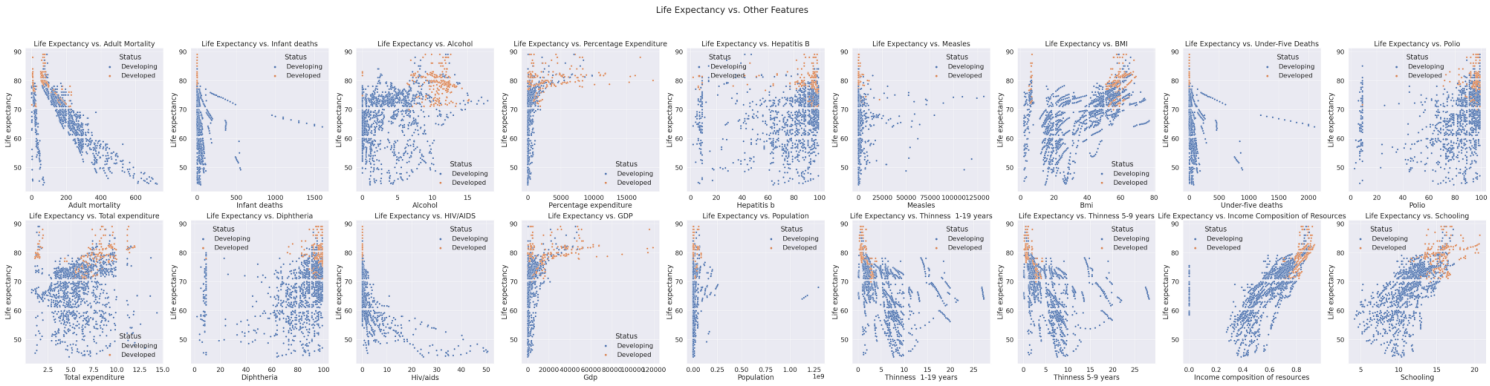

Figure 3: Scatter Plots of Life Expectancy vs. Other Characteristics

It seemed like life expectancy was affected by all the factors that were present in the data set. I created a scatter plot to confirm that there was a relationship between life expectancy and the other variables. The results that I got from the correlation matrix relating to Life Expectancy coincided with the scatterplot results specifically Adult Mortality, Income Composition of Resources, and Schooling. Life expectancy works the best out of all of the variables as the target variable because the other variables play a role in determining it.

Choosing Models

Since the target variable, Life Expectancy, is a continuous variable, I chose machine learning models that fitted well for a regression problem. The regression models I used include a linear regression model, random forest regression model, decision tree regression model, and lastly an ada boost regression model. For all of the models, my target variable was life expectancy and my predicting variables were all of the other columns in my dataset. I made status into a dummy variable. I used a test size of 0.3 for all of my models, fit the training data into the model, and made the model predict the X_test value.

Training/Testing Dataset Results for Models

Linear Regression Model

After creating the linear regression model, I looked into the r-squared values for the training set and the testing set. The training set had an r2 score of 0.834 and the testing set had an r2 score of 0.835. The residuals plot was pretty symmetrically distributed, there weren't any clear patterns, and the points were clustered around 0. The distribution of the residuals was centered at 0 and normal.

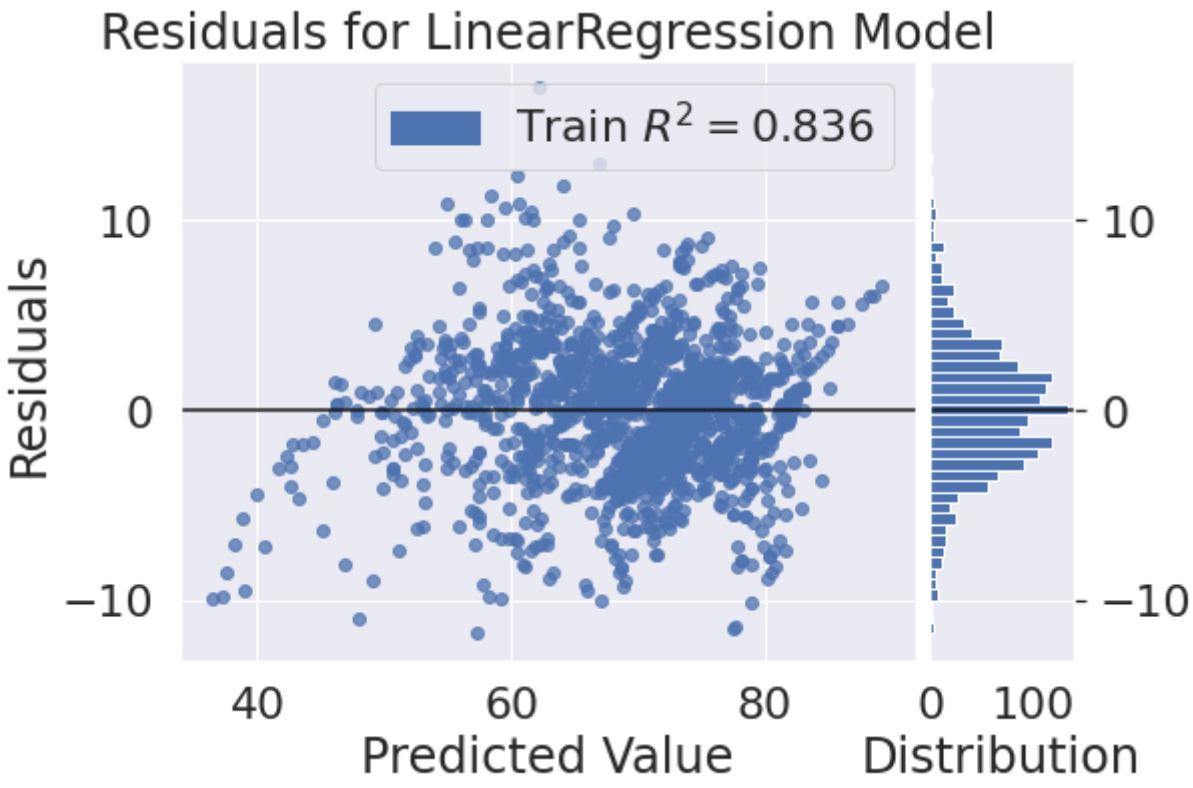

Figure 5: Residuals Plot for Linear Regression Model

Random Forest Model

I proceeded to follow the same steps that I did for the Linear Regression model for my Random Forest model. The training set had an r2 score of 0.993 and a testing r2 score of 0.963. The mean square error value was 1.73. Since the value was low, the model was better at predicting accurately.

Decision Tree Model

For this model, the training set had an r2 score of 0.845 and a testing r2 score of 0.827. The training r2 score was better than the linear regression model but worse than the random forest model. For the testing r2 score, it did worse than the other two models.

ADA Boost Regression Model

For this model, the training set had an r2 score of 0.853 and a testing r2 score of 0.804. It had a better training r2 score than the decision tree model but had a lower r2 testing score than the rest of the other models.

Choosing the Most Promising Model

The **Random Forest Model** scored the best out of all of the models for both the training data score and the testing data score. Therefore, I picked this model to fine-tune the model's parameters and improve its performance.

Cross-Validation to Fine Tune Parameters

To cross-validate my model, I used the Randomized Search CV method first and then the Grid Search CV to build off of that. Both methods further split the training data into subsets called folds. The model trains the data on K-1 of the folds and evaluates (validates) the Kth fold. Through the Randomized Search CV, I was able to define a grid of hyperparameter ranges and perform K-fold CV. The base model had an average error of 1.2436 degrees and an accuracy of 98.16% which was already very high. However, I was able to get a 0.16% improvement in my model by reducing the error to 1.1242 degrees and improving the accuracy to 98.32% with the Randomized Search CV.

I continued to improve my model using Grid Search CV that evaluates combinations based on what I define. Through this method the average error got further reduced to

0.5332 degrees and had an accuracy of 99.2%, creating an overall improvement of 1.05% which was a lot larger than the previous Randomized Search CV method used. The r-square value for predicting the life expectancy was 0.99.

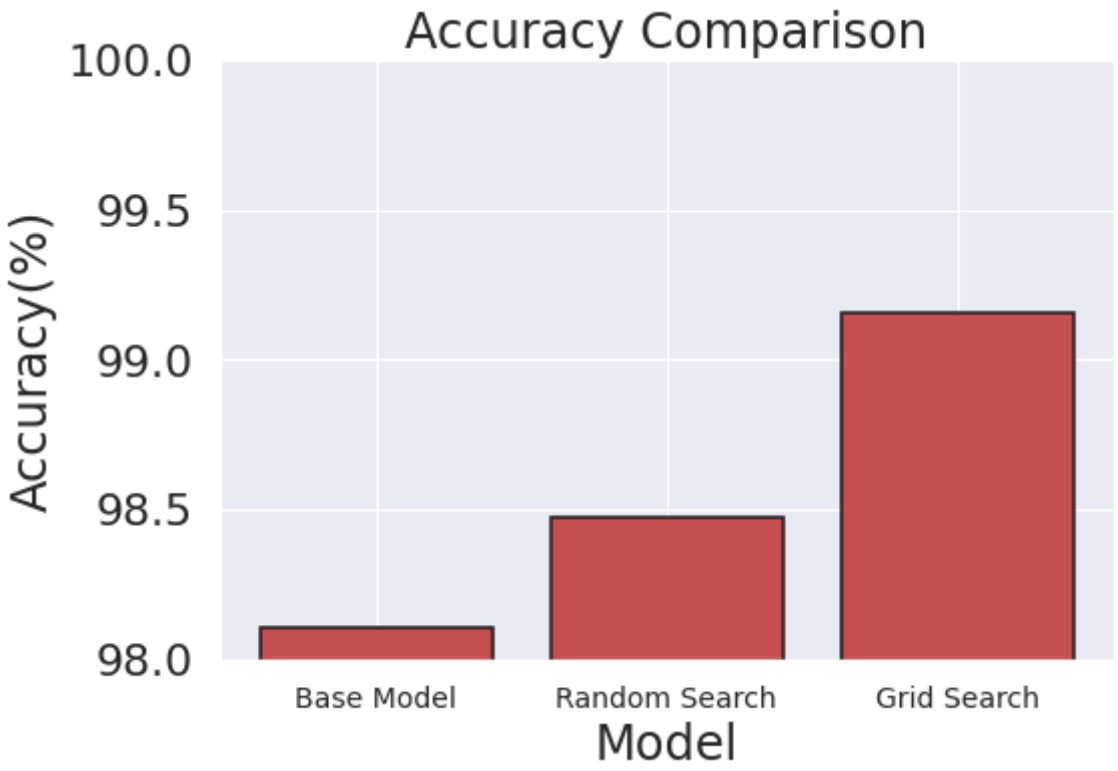

Figure 6: Bar Chart Comparing Accuracies for the Fine-Tuning

After fine-tuning the model, I looked at the residuals that displayed the difference between the predictions made by the model and the actual results of the test data. The distribution of the residuals was normal with a mean-centered around 0. This shows that my model was highly efficient.

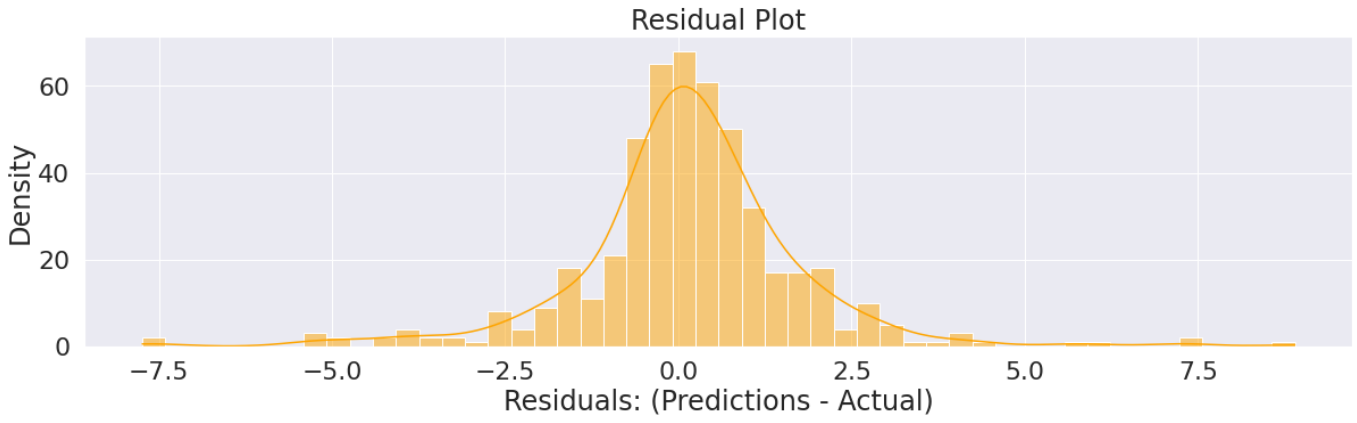

Figure 7: Residuals Plot After Fine-Tuning

Simplifying the Model

I was worried that my model might be overfitting the training data, so I simplified it by using a regularization method called Ridge Regression. I calculated the mean absolute error after getting the cross-validation scores using the Ridge Regression and I got a value of 2.779. The low value confirms the good predicting ability of the model.

Machine Learning